复杂大交通场景弱小目标检测技术 *

华 夏 [1]，王新晴 [1]，马昭烨 [1]，王 东 [1,2]，邵发明 [1]

(1. 陆军工程大学，南京 210007; 2. 南部战区陆军第二工程科研设计所，昆明 650222)

摘 要：针对现有基于大数据和深度学习的目标检测框架对于高分辨率复杂大场景中低分辨率小目标识别效果较差，多目标检测的精度和实时性难以平衡的问题,改进了基于深度学习的目标检测框架 SSD(single shot multibox detector)，提出一种改进的多目标检测框架 DRZ-SSD（DRZ），将其专用于复杂大交通场景多目标检测。检测以从粗到细的策略进行，分别训练一个低分辨率粗略检测器和一个高分辨率精细检测器，对高分辨率图像进行下采样获得低分辨率版本，设计了一种基于增强学习的动态区域放大网络框架（DRZN），动态放大低分辨率弱小目标区域至高分辨率再使用精细检测器进行检测识别，剩余图像区域使用粗略检测器进行检测，对弱小目标的检测与识别精度以及运算效率的提高效果明显；采用模糊阈值法调整自适应阈值策略在避免适应数据集的同时提高模型的决策能力，显著降低检测漏警率和虚警率。实验表明，改进后的 DRZ-SSD 在应对弱小目标、多目标、杂乱背景、遮挡等检测难度较大的情况时，均能获得较好的效果。通过在指定数据集上测试，相比于其他基于深度学习的目标检测框架，各类目标识别的平均准确率提高了 4~15%，平均准确率均值提高了约 9~16%，多目标检测率提高 13~34%，检测识别速率达到 38 帧/s，实现了算法精度与运行速率的平衡。

关键词：机器视觉；深度学习；神经网络；交通场景多目标检测；增强学习；自适应

中图分类号：TP391.4 doi: 10.3969/j.issn.1001-3695.2018.05.0343

Detection of dim and small targets in complex large traffic scenes

Hua Xia[1], Wang Xinqing[1], Ma Zhaoye[1], Wang Dong[1, 2], Shao Faming[1]

(1. *Army Engineering University*, *Nanjing* 210007, *China*; 2 *Second Institute of Engineering Research & Design*, *Southern Theatre Command*, *Kunming* 650222, *China*)

Abstract: Aiming at the problems that the existing target detection framework based on big data and depth learning has poor recognition effect on low-resolution small targets in high-resolution complex large-field scenes, and the accuracy and real-time performance of multi-target detection are difficult to balance, improve the single shot multi-box detector based on depth learning, and propose an improved multi-target detection framework DRZ - SSD (dynamic region zoom - in, DRZ) , which is dedicated to multi-target detection in complex large traffic scenes. The detection is carried out in a coarse-to-fine strategy, training a low-resolution coarse detector and a high-resolution fine detector respectively, downsampling the high-resolution image to obtain a low-resolution version, designing a dynamic region zoom - in network based on enhanced learning, dynamically enlarging the low-resolution small target region to a high-resolution and then using the fine detector to carry out detection and identification, and detecting the remaining image region by using the coarse detector, so that the detection and identification accuracy of the small target and the improvement effect of the operation efficiency are obvious; Adopting fuzzy threshold method to adjust the adaptive threshold strategy can not only avoid adapting to the data set but also improve the decision-making ability of the model and significantly reduce the detection missed alarm rate and false alarm rate. Experiments show that the improved drz - SSD can achieve good results when dealing with weak targets, multi - targets, cluttered background, occlusion and other difficult detection situations. Through testing on the specified data set, compared with other target detection frameworks based on in-depth learning, the average accuracy rate of various types of target recognition has increased by 4~15 %, the average accuracy rate has increased by 9~16 %, the multi-target detection rate has increased by 13~34 %, and the detection and recognition rate has reached 38 frames / s, realizing the balance between the accuracy of the algorithm and the running rate.

Key words: machine vision; deep learning; neural network; traffic scene multi-target detection; reinforcement learning; self-adaptation

0 引言

交通场景中的行人、车辆目标检测与识别是目标检测技术的重要分支，是自动驾驶、机器人以及智能视频监控等研究领域的核心技术，有着重要的研究意义[1]。

深度学习为基于深层人工神经网络的学习方法，基于深度学习的目标检测算法可应用于多种检测场景，综合性强，能够同时检测和识别多类目标，主动性好。各种类型的人工神经网

收稿日期：2018-05-23；**修回日期：**2018-07-30 **基金项目：**国家重点研发计划资助项目；国家自然科学基金资助项目；江苏省自然科学基金资助项目；中国博士后科学基金第 62 批面上资助项目）

作者简介：华夏（1995-），男，硕士研究生，主要研究方向为计算机图形学、机器视觉、数字图像处理（1614118084@qq.com）；王新晴（1963-），男，教授，博导，博士，主要研究方向为机电控制、智能信号处理、机器视觉；马昭烨，男，讲师，主要研究方向为机电控制、智能信号处理、机器视觉；王东，男，讲师，博士，主要研究方向为机电控制、智能信号处理、机器视觉；邵发明，男，讲师，博士研究生；主要研究方向为机电控制、智能信号处理、机器视觉.

络结构中，深度卷积网络具有强大的特征提取能力，越来越多的用于图像分类的网络结构被提出，不断提升了深度卷积网络在特征提取方面的优势，在图像识别、图像分割、目标检测、场景分类等视觉任务中，取得了非常好的效果[2]。Faster RCNN[4]替代掉费时的 selective search 方法，速度提高了，RPN 产生的 region proposal 质量高，准确率（mAP）也提高了，但是 NPR 产生的在图像边缘的 region proposal 信息被丢弃了。YOLO[5]将物体检测作为回归问题进行求解，整个检测网络 pipeline 简单，且训练只需一次完成, YOLO 在训练和推理过程中能“看到”整张图像的整体信息，背景误检率低,而基于 region proposal 的物体检测方法（如 Fast RCNN）在检测过程中，只“看到”候选框内的局部图像信息,但是识别物体位置精准性差，召回率低，尤其是对小目标和密集目标检测识别效果差。

SSD，全称 single shot multibox detector[3]，是 Liu Wei 在 ECCV 2016 上提出的一种目标检测算法，截至目前是主要的检测框架之一，相比 Faster RCNN[4]有明显的速度优势，相比 YOLO[5]又有明显的平均准确率均值（mAP）优势。SSD 具有如下主要特点：从 YOLO 中继承了将 detection 转化为 regression 的思路，同时一次即可完成网络训练；基于 Faster RCNN 中的 anchor，提出了相似的 prior box；加入基于特征金字塔[6]（Pyramidal Feature Hierarchy）的检测方式，相当于半个 FPN[6]思路。尽管 SSD 在特定数据集上已经取得了较高的准确率，具有较好的实时性，但是模型的训练过程非常耗时，对训练样本的质和量依赖严重；通过图像的颜色、边缘等信息来检测目标，其对于弱小目标和大面积遮挡目标等缺乏图像信息的目标检测效果不佳；算法检测效率仍然有待提高，以满足装备运行实时性的要求。

本文针对复杂大交通场景下行人、车辆目标检测任务的特点和需求，对传统 SSD 算法进行了以下两点改进：1）利用增强学习和顺序搜索方法，结合大交通场景目标检测任务的特点和需求，提出了一种动态区域放大网络框架（Dynamic Region Zoom-in Network,DRZN），该网络框架通过下采样图像，大幅降低了运算量，同时通过动态区域放大保持了高分辨率图像中不同尺寸目标的检测精度，对低分辨率弱小目标的检测与识别精度提高效果明显，降低检测漏警率；2）针对 SSD 检测固定置信度阈值不够灵活的缺陷，采用模糊阈值法调整自适应阈值策略在避免适应数据集的同时提高模型的决策能力，显著降低检测漏警率和虚警率。

1　动态区域放大网络框架

SSD 采用了特征金字塔结构进行检测，即检测时利用了 conv4-3，conv-7（FC7），conv6-2，conv7-2，conv8_2，conv9_2 这些大小不同的 feature maps，在多个 feature maps 上同时进行 softmax 分类和位置回归，对弱小目标有较好的检测精度[3]，但是在复杂大交通场景下对低分辨率弱小目标的检测效果仍然不够理想。

针对 SSD 存在的复杂大场景下对于低分辨率弱小目标检测困难问题，本文提出了一种动态区域放大网络框架（DRZN），该网络框架通过对高分辨率大场景图像进行下采样，降低了目标检测的计算量，同时通过动态区域放大保持了高分辨率图像中低分辨率弱小目标的检测精度，对弱小目标的检测与识别精度的提高效果明显。检测以从粗到细的方式进行，首先对图像的下采样版本进行检测，然后对被识别为可能提高检测精度的区域顺序放大至较高分辨率版本再进行检测。该方法建立在增强学习的基础上，由一个放大精度增益回归网络[7](R-net)和一个放大区域动态选择算法（Zoom-in Region Choose）两部分组成，前者学习粗检测和精检测之间的相关性，并预测放大区域的精度增益，后者依据前者学习和预测结果动态选择需要被放大的区域。

首先对图像的下采样版本执行粗略检测，以降低运算量提高运行效率，然后顺序地选择可能存在低分辨率小目标的区域进行放大操作然后分析来保证对低分辨率小目标的识别精度。采用强化学习方法从检测精度和计算成本两个方面对放大奖励进行建模，并动态选择一系列区域放大至高分辨率再进行分析。算法总体框架如图 1 所示。

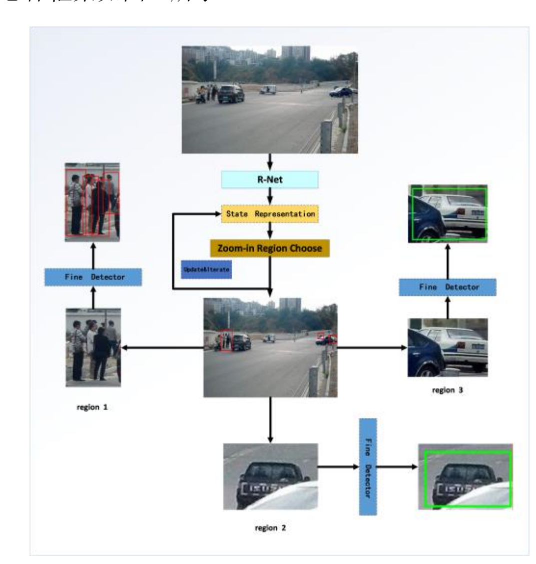


图 1　动态区域放大网络架构

1.1　放大精度增益回归网络 R-net

顺序搜索。处理高分辨率大场景图像的策略是避免处理整个图像，而是顺序地检测疑似目标的小区域。

强化学习（reinforcement learning ,RL）。RL 是用于学习顺序搜索策略的通用机制，因为它允许模型考虑一系列动作的效果而不仅仅是单个动作的效果[8]。RL 是在尝试的过程中学习到在特定的情境下选择哪种行动可以得到最大的回报。在很多场景中，当前的行动不仅会影响当前的收益，还会影响之后的状态和一系列的收益。RL 最重要的三个特征在于：基本是以一种闭环的形式；不会直接指示选择哪种行动；一系列的行动和奖励信号对之后的行动都会产生较长时间的影响。RL 采用的是边获得样本边学习的方式，在获得样本之后更新自己的学习

模型，利用当前的模型来指导下一步的行动，下一步的行动获得收益回馈之后再更新学习模型，不断迭代重复直到学习模型达到收敛。

本算法采用由粗到细的检测策略，在低分辨率下应用粗检测器，并利用该检测器的输出结果来指导对高分辨率目标的深入搜索。虽然粗略检测器将不如精细检测器精确，但它将识别需要进一步分析的图像区域，从而仅在有希望的区域中产生高分辨率检测的运算成本。算法主要运用由两个机制组成：a）学习粗检测器和细检测器之间的统计关系的机制，以便在给定粗检测器输出的情况下预测哪些区域需要放大；b）用于在给定粗略检测器输出和需要由精细检测器分析的区域的情况下选择要以高分辨率分析区域的序列的机制。

本文策略可以被表述为马尔可夫决策过程。在每个步骤，系统先观察当前状态，估计采取不同行动的潜在成本感知回报，并选择具有最大长期成本感知回报的行动[9]。要素包括:

a）动作。该算法以高分辨率依次分析具有高放大回报的区域。在此上下文中，动作对应于选择要以高分辨率分析的区域。每个动作可以由向量(x,y,w,h)来表示；其中(x,y)表示指定区域位置，(w,h)表示指定区域的大小。在每个步骤中，该算法根据潜在的长期奖励对采取一组潜在的动作(矩形区域的列表)进行评分。

b)状态集。表示编码两种类型的信息:待分析区域的预测精度增益；以及已经以高分辨率分析的区域的历史(同一区域不应被多次放大)。本文设计了一个放大精度增益回归网络(R-net)来学习信息精度增益图(AG map)作为状态表示。AG map 具有与输入图像相同的宽度和高度。AG map 中的每个像素的值是对输入图像中包括那个像素可以提高多少检测精度的估计。所以，AG map 提供了用于选择不同动作的检测精度增益。在采取动作之后，对应于 AG 映射中所选区域的值相应地减小，因此 AG 映射可以动态地记录动作历史。

c)损失回报函数。状态对放大每个图像子区域的预测精度增益进行编码。为了在有限的计算量下保持高精度，定义了一个损失回报函数如式 1。给定状态和动作，损失回报函数通过考虑成本增量和精度改进对每个动作(缩放区域)进行评分

$$R\left(s_{tates},a_{ctions}\right)=\sum_{k\in a_{ctions}}\left|g_k-p_k^l\right|-\left|g_k-p_k^h\right|-\lambda_2\frac{b_1}{B} \tag{1}$$

其中：动作中 k 表示目标 k 包含在由动作选择的区域中。p_k^l 和 p_k^h 表示对同一目标粗略检测器和精细检测器的目标检测分数，且 g_k 是对应的目标真实标签。变量 b_1 表示所选区域中的像素的总数，B 表示输入图像的像素的总数。式中第一项表示检测精度的提高。第二项表示放大成本。精度和计算之间的平衡由参数 λ_2 控制。

放大精度增益回归网络（R-Net）基于粗略检测结果预测特定区域上放大的精度增益。R-Net 在粗检测和精检测数据对上训练，以便它可以观察它们如何相互关联以学习适当的精度增益关系[7]。

由于 SSD 在许多计算机视觉应用中的成功，使用 SSD 作为基础检测器。两个 SSD 分别在高分辨率精细图像组成的训练集和低分辨率粗略图像组成的训练集上进行训练，并随后用作黑盒粗略和精细检测器。将两个预先训练好的检测器应用于一组训练图像并获得两组图像检测结果：下采样图像中的低分辨率检测$\left\{\left(d_i^l,p_i^l,f_i^l\right)\right\}$和在每个图像的高分辨率版本中的高分辨率检测$\left\{\left(d_j^h,p_j^h\right)\right\}$其中 d 是检测边界框，p 是作为目标对象的概率，f 表示相应检测的特征向量。使用上标 h（High）和 l（Low）来表示高分辨率和低分辨率（下采样）图像。

为了使模型判别高分辨率检测是否改善了整体检测结果，引入了一个匹配层，将两个检测器产生的检测结果关联起来。在该层中，如果发现下采样图像中的可能对象 i 和高分辨率图像中的可能对象 j 具有足够大的交集 $IoU\left(d_i^l,d_j^h\right)$（$IoU>0.5$），则定义 i 和 j 为彼此对应。按照规则对粗检测方案和精检测方案进行匹配，并生成它们之间的一组对应关系[7]。

给定一组对应关系$\left\{\left(d_k^l,p_k^l,p_k^h,f_k^l\right)\right\}$，可以估计粗检测的放大精度增益。检测器只能处理一定范围内的对象，因此将检测器应用于高分辨率图像并不总是产生最佳精度。例如，如果检测器主要在小目标数据集上训练，则该检测器对较大目标的检测精度并不高。因此，使用$\left|g_k-p_k^l\right|-\left|g_k-p_k^h\right|$来测量哪个检测结果(粗略或精细)更接近事实，其中 $g_k\in\{0,1\}$ 作为真实标签的度量。当高分辨率分数 p_k^h 比低分辨率分数 p_k^l 更接近基本事实时，该函数表示此目标值得放大；否则，在下采样图像上应用粗略检测器可能产生更高的精度，因此我们应该避免放大该目标。使用相关回归(CR)层来估计目标 K 的放大精度增益

$$\min_{W_1}\left(\left|g_k-p_k^l\right|-\left|g_k-p_k^h\right|-\phi\left(W_1,f_k^l\right)\right)^2 \tag{2}$$

其中：∅代表回归函数，W_1 代表参数集。该层的输出是估计的准确度增益。CR 层包含两个完全连接的层，第一层有 4096 个单元，第二层只有一个输出单元。

根据每个目标的学习准确性增益可以生成 AG map（accuracy gain map）。假设候选边框内的每个像素对其准确性增益具有同等的贡献。因此，AG map 生成为

$$AG(x,y)=\begin{cases}\alpha\dfrac{\phi\left(\hat{W},f_k^l\right)}{b_k} & if\ (x,y)\ in\ d_k^l\\ 0 & otherwise\end{cases} \tag{3}$$

其中：(x,y) 表示点(x,y)在边界框 d_k^l 内，b_k 表示包含在

d_k^l 中的像素数。α 是一个常数。$\hat{W}$ 表示 CR 层的估计参数。

AG map 用作状态表示，它自然包含粗略检测质量的信息。 在对区域进行放大和检测后，区域内的所有值均设置为 0，以防止未来在同一区域再次进行缩放。放大精度增益回归网络 R-net 结构图如图 2 所示

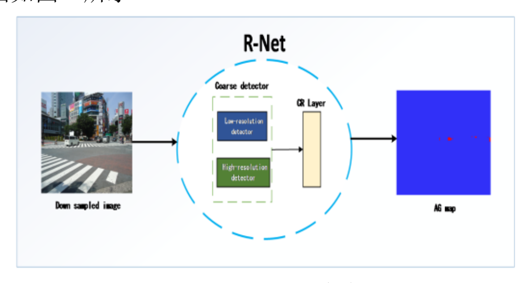


图 2　R-Net 网络框架

1.2　放大区域动态选择算法

通过 R-net 获得了 AG map，AG map 中的每个像素的值是对输入图像中包括那个像素可以提高多少检测精度的估计。所以，AG map 提供了用于选择不同动作的检测精度增益。在采取动作之后，对应于 AG 映射中所选区域的值相应地减小，因此 AG 映射可以动态地记录动作历史。依据 AG map 提出了一种动态放大区域选择算法，具体算法流程如图 3 所示。

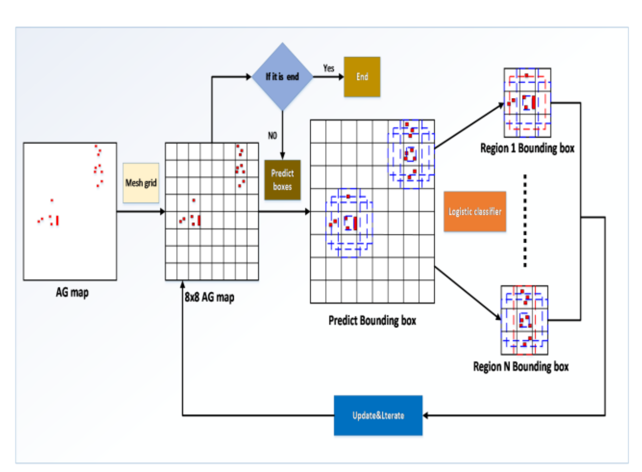


图 3　动态放大区域选择流程

首先将 AG map 按照 8x8 网格划分为等额矩形区域，统计每个矩形中像素值的总和，设定阈值，选择区域中心块，每个区域中心块为中心的 3x3 个矩形构成放大筛选区域，同一个放大筛选区域类如有多个满足像素值阈值条件的矩形区域，取像素值最大的那个作为区域中心，如果区域中心取在大正方形的边上，通过增补同尺寸空白小正方形的方式构成 3x3 的放大筛选区域。在放大筛选区域内，以放大筛选区域中心点为中心，按照不同的长宽比，构造 4 个不同长宽比的预测包围盒，通过比较各个预测包围盒包含区域的构造指标（像素值、比例、面积）选出最佳放大区域包围盒。网格划分后的 AG map 中矩形区域内 rtg_i 的总像素值 $s_{ump}x_i$

$$s_{ump}x_i = \sum_{j\in rtg_i} px_j \tag{4}$$

其中 px_j 代表 rtg_i 区域内第 j 个像素点的像素值，$s_{ump}x_i$ 值越大，代表矩形区域 rtg_i 的放大收益越大，将高放大收益的区域作为中心符合人眼对区块领域相关性的认识。通过二阶差分法自适应选取像素值阈值，完成区域中心块的初筛选。二阶差分可以表现离散数组的变化趋势大小，可用于在一组像素值中确定阈值。检测一张 AG map 默认得到 64 个候选区域，

最后每个候选区域都得到 1 个用来表示放大收益的总体像素值 $s_{ump}x_i$，故共可以得到的 64x1 的数组，舍去其中小于 0.1 的元素，判为没有目标，得到 $n\times1$ 的数组 C。设估计 $s_{ump}x_i$ 由大减小变化趋势的函数为 $f(g)$，见式 5

$$f(C_k)=\frac{(C_{k+1}-C_k)-(C_k-C_{k-1})}{C_k},k=2,3,\cdots,n-1 \tag{5}$$

则将 $f(C_k)$取最大值时的 C_k 作为此 AG map 图像的 $s_{ump}x_i$ 阈值。

为了减少区域放大精检测的计算量，有效提高算法的效率和实时性，同时又要保证所选区域有较好的包容度，以每个区域中心块为中心的 3x3 个矩形构成放大筛选区域，同一个放大筛选区域类如果有多个满足像素值阈值条件的矩形区域，取像素值最大的那个作为区域中心。

以放大筛选区域中心点为中心位置，按照不同的长宽比预测 6 个固定大小的预测包围盒，放大筛选区域的面积为 S_Z 每个预测包围盒的面积如式（6）所示

$$s_k = s_{min} + \frac{s_{max}-s_{min}}{m-1}(k-1)\ \ ,k=1,2,\cdots,5 \tag{6}$$

其中 $s_{min}=0.1\times S_Z$，$s_{max}=0.7\times S_Z$，$m=5$，对于不同的预测包围盒赋予不同的长宽比：

$$a_r=\frac{W}{H},\ a_r\in\left\{1,2,3,\frac{1}{2},\frac{1}{3}\right\} \tag{7}$$

W、H 分别表示包围盒的宽和长。则预测包围盒对应的宽和长分别为 $H_k=\sqrt{s_k/a_r}$，$W_k=\sqrt{a_r\bullet s_k}$。

当 $a_r=1$ 时还有一个预测包围盒，规模为 $s_k^{'}=\sqrt{s_k\bullet s_{k+1}}$，即一共有 6 个预测包围盒。

对于任一个包围盒，b_l，计算盒内的像素总值 $s_{ump}x_i$ 为

$$s_{ump}x(b_l)=\sum_{i\in b_l} px_i\ ,\ l=1,2,3,4 \tag{8}$$

区域面积 S 为

$$S(b_l)=W\times L \tag{9}$$

W、H 分别表示盒的宽和长。区域内高放大收益像素占比 P 为

$$P(b_l)=\frac{pn_1}{pn_2} \tag{10}$$

Pn_1 表示 b_l 区域内，具有放大收益的像素点（即像素值大于 0.1 的像素点）的总数，Pn_1 表示 b_l 区域像素点总数。即每一个预测包围盒，b_l 存在特征向量$(x,y,sumpx,W,L,P)$，x、y 分别

表示 b_l 的中心点横纵坐标。

利用人工标定的训练样本，训练了一个 Logistic 分类器[10]对各个预测包围盒的框选效果进行评价。将评价结果分为两类，即能够满足放大要求的预测包围盒和不能满足放大要求的预测包围盒。

对于输入的预测包围盒，$b_l\,(x,y,sumpx,W,L,P)$，Logistic 分类器引入权值参数 $\theta=(\theta_1,\theta_2,K,\ \theta_6)$ ，对 b_l 中的属性进行加权，得到 $\theta^T b_l$ ；引入 logistic 函数（sigmoid 函数）得到函数 $h_\theta(b_l)$

$$h_\theta(b_l)=\frac{1}{1+e^{-\theta^T b_l}} \tag{11}$$

即可得到概率估计函数 $P(y\mid b_l;\ \theta)$

$$P(y\mid b_l;\ \theta)=\begin{cases}h_\theta(b_l);\ y=1\\1-h_\theta(b_l);\ y=0\end{cases} \tag{12}$$

它的含义就是在给定测试样本 b_l 与参数 θ 时，标签为 y 的概率。

由测试样本集合与训练样本集合，我们可以得到它们的联合概率密度即似然函数：

$$\prod_{i=1}^{n}P(y^{(i)}\mid b_l^{(i)};\ \theta)=\prod_{i=1}^{n}\left(h_\theta(b_l)\right)^{y(i)}\left(1-h_\theta(b_l)\right)^{1-y(i)} \tag{13}$$

最大化似然函数，求出合适的参数 θ。将式 13 变形为

$$\ell(\theta)=\sum_{i=1}^{m}y^{(i)}\log h_\theta(b_l)+\left(1-y^{(i)}\right)\log\left(1-h_\theta(b_l)\right) \tag{14}$$

依据公式，由梯度下降法求取参数 θ。先对参数 θ 求导

$$\frac{\partial}{\partial\theta_j}\ell(\theta)=\left(y-h_\theta(b_l)\right)b_{l\,j} \tag{15}$$

更新法则

$$\theta_j:=\theta_j+\alpha\left(\left(y^{(i)}-h_\theta\left(b_l^{(i)}\right)\right)b_{l\,j}^{\ (i)}\right) \tag{16}$$

通过 Logistic 分类器对各个预测包围盒的框选效果进行评价后，对于每一个预测包围盒我们都能够获得一个对应的框选评价分数，之后，进行一个非极大值抑制得到最终的预测作为最终的放大包围盒。

在完成放大包围盒的选取后我们将放大筛选区域内的像素值全部设为 0，避免重复选取造成的效率低下，同时对 AG map 进行对应区域的更新，并检测 AG map 上是否已经对所有高放大收益区域进行检测（AG map 像素总值是否为 0），若果是则完成检测，否的话继续迭代进行检测过程。

把所得放大精检测候选区域的原图部分送到精细检测器检测之前，先进行双线性插值放大，放大至精细检测器检测候选区域的最小尺寸（本文设置的候选区域最小为 10×10）。

2　置信度自适应阈值改进

在 SSD 用 Softmax 为候选区域进行分类的最后阶段，候选区域会得到属于各个类别的置信度（即属于各个类别的概率），当属于某类的置信度高于设定阈值时则将此候选区域判为该类目标，若同一候选区域有多个类别置信度高于阈值则取最高者。

目标尺度较小或被遮挡时置信度相对较低。若采用固定阈值，设置过高会排除许多真目标，过低会混入许多假目标。通常的做法是不断调整阈值对数据集进行多次测试，计算出不同阈值下的平均准确率，取平均准确率最大的阈值作为模型最终的阈值。但这种做法有适应数据集的倾向，再庞大的数据集也无法涵盖现实中的所有情况。本文采用自适应阈值在避免适应数据集的同时提高模型的决策能力[11]。

无论固定或自适应，阈值的设定都需要参考数据集中目标得分情况。检测模型训练较好的情况下，正确的检测结果中真目标和假目标置信度常常相差一两个数量级，且真目标置信度通常在 0.7 以上。虽然与真目标的置信度存在差距，但假目标也会因为某些特征与目标类似而取得 0.7 以上的高置信度，单纯采用固定阈值无法将目标与背景区分开[11]。针对 SSD 检测固定置信度阈值不够灵活的缺陷，采用模糊自适应阈值法[12]调整自适应阈值策略降低漏警率和虚警率。

模糊程度是由模糊率函数来确定，当模糊率最低的时候，这时候分割效果最好。其中模糊率与隶属函数相关，模糊数学的基本思想是隶属度的思想。应用模糊数学方法建立数学模型的关键是建立符合实际的隶属函数[12]。

检测一张图像默认得到 N 个候选区域送入 SSD，最后每个候选区域都得到 M 个用来表示属于 M 个类别的置信度，故共可以得到的 N 个 $M\times1$ 的数组。取出每个数组中的最大值并由大到小排序，舍去其中小于 0.1 的值(若 N 个值全部小于 0.1 则判为没有目标)，得到 $N\times1$ 的数组 C。$\mu(x)$ 是隶属度函数，$\mu(C_k)$ 为数组 C 中置信度取 Ck 的区域的隶属度。数组 C 的模糊率 $\gamma(C)$ 是对数组 C 的模糊性度量，令 $h(C_k)$ 为数组 C 中置信度取 Ck 的元素个数，则数组 C 的模糊率 $\gamma(C)$ 定义如式 17

$$\gamma(C)=\frac{2}{n}\sum_{k=0}^{n-1}T(C_k)h(C_k) \tag{17}$$

其中 $T(C_k)=\min\{\mu(C_k),1-\mu(C_k)\}$ 。

数组 C 的模糊率 $\gamma(C)$ 取决于隶属度函数 $\mu(x)$，若取隶属度函数为 S 函数，即

$$\mu(x)=\begin{cases}0, & 0\le x\le q-\Delta q\\2\left[\dfrac{(x-q+\Delta q)}{2\Delta q}\right]^2, & q-\Delta q\le x\le q\\1-2\left[\dfrac{(x-q+\Delta q)}{2\Delta q}\right]^2, & q<x\le q+\Delta q\\1, & q+\Delta q<x\le C_n\end{cases} \tag{18}$$

则此时 $\mu(x)$ 由窗宽 $c=2\Delta q$ 及参数 q 决定，一旦选定了窗宽，则 $\gamma(C)$ 就只与参数 q 有关。模糊阈值法的求解过程是预先设定窗宽，根据论文前人的研究，系数常设定为 0.3。通过改变 q 使得隶属度函数 $\mu(x)$ 在置信度区间 $[C_0,C_{n-1}]$ 上滑动，通过计算模糊率 $\gamma_q(C)$ 获得模糊率曲线，该曲线的谷点，即使 $\gamma_q(C)$ 取得极小值的 q，也就是所求的自适应阈值。

改进后整体的检测算法框架流程如图 4 所示

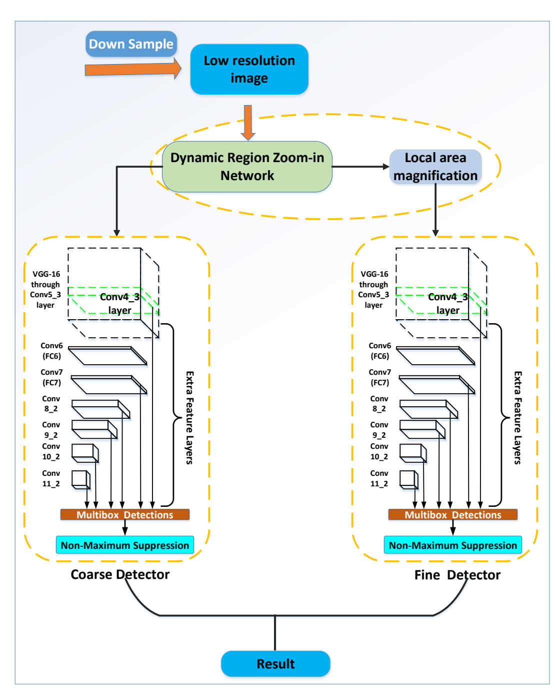


图 4　改进后算法整体框架

算法流程如下：a)输入待检测视频单帧图像，将图像进行下采样获得低分辨率版本，降低运算量；b)通过 DRZN 对需要在高分辨率下检测识别的目标区域进行顺序选择放大，精细检测器进行目标检测识别，获得结果 R_1，已经检测过的区域在原低分辨率图像中像素值全取为 0；c)将剩余低分辨率图像输入粗略检测器进行目标检测识别，获得结果 R_2；d)将 R_1 和 R_2 检测结果进行融合，得到最终结果 R_3。

3　实验结果与分析

3.1　实验的基础条件与数据集库

本文实验使用 DELL Precision R7910(AWR7910) 图形工作站，处理器为 Intel Xeon E5-2603 v2(1.8 GHz/10M)，采用 NVIDIA Quadro K620 GPU 加速运算。SSD 是基于深度学习框架 Caffe 来运行的。Caffe 支持 CPU 和 GPU 的并行运算，使得计算量庞大的深度学习得以在短期内完成。

本文在 YFCC100M 收集的交通场景数据集(Web dataset)和 KITTI 数据集上进行了实验。选用 KITTI 数据集中第 1 个图片集“Download left color images of object data set”和标注文件“Download training labels of object data set”，其中 7 481 张训练图片有标注信息，而测试图片没有。SSD 中训练脚本是基于 VOC 数据集格式的，需要把 KITTI 数据集做成 Pascal VOC 的格式。Pascal VOC 数据集总共 20 个类别，本文为数据集设置 3 个类别‘Car’，‘Cyclist’，‘Pedestrian’，因为标注信息中还有其他类型的车和人，本文将 ‘Van’‘Truck’‘Tram’ 合并到 ‘Car’ 类别中去，将‘Person_sitting’合并到‘Pedestrian’类别中去，‘Misc’ 和 ‘Dontcare’ 这两类直接忽略。在测试集中选出 100 张含低分辨率小目标（本文设定小目标尺寸小于 10×10 的图像），构成 KITTI 数据集的低分辨率小目标测试集。

YFCC100M 数据集包含将近 1 亿张图片以及摘要、标题和标签。为了更好地展示本文方法，从 YFCC100M 数据集收集了 1000 幅分辨率较高的测试图像。通过搜索关键词“行人”、“道路”和“车辆”来收集图像。对于该数据集，使用至少 16 像素宽度和小于 50 %遮挡对所有目标进行注释。图像在较长的一侧被重新缩放到 2000 像素，以适合 GPU 内存。在测试集中选出 100 张含低分辨率小目标（本文设定小目标尺寸小于 10×10 的图像），构成 WD 数据集的低分辨率小目标测试集。实验中将所有的图像尺寸归一化为 320 x 320。

3.2　实验的参数设置

本文选择 SSD 系列中的 SSD512 进行改进，SSD512 提供了大、中、小三个规模的深度卷积神经网络模型，本文选取中等规模的 VGG_CNN_M_1024 模型作为基础模型，改动与目标类别数目相关的参数（原模型需要识别 20 类目标而本文只有 3 类）。小样本数据集在一定的程度上可以代表原始数据集，通过小样本数据集训练所得的最优超参数在一定的程度上能够适应原始数据集[13]。通过小样本调参，在不使用自适应阈值时，阈值设置为 0.1（默认设置为 0.7）；将所有实验中经过非极大抑制留下的候选区域数量设置为 100（默认设置为 300）。其他设置保持默认不变，后续所有实验都在以上设置基础上进行。

3.3　评价指标

在多目标分类器的判别中，设目标的种类数为 n 。对单种目标的判别仍然遵循每一种假设有两种结果的四种可能性，即

设 $D_i^j\left(j=1,2,\cdots,n\right)$ 表示一种目标 j 选择假设 H_i^j 为真，任何二元假设实验问题中， 作判别时要考虑 4 种可能性[14]：

a）H_0^j 假设为真，判别为 D_0^j；b）H_0^j 假设为真，判别为 D_1^j；c）H_1^j 假设为真，判别为 D_0^j；d）H_1^j 假设为真，判别为 D_1^j。

a）和 d）对目标 j 选择正确；b）称为第一类错误，叫做虚警（没有目标而识别为有目标）；c）称为第二类错误， 叫做漏报（有目标而误判为没有目标）。除此之外，在多目标识别中将目标 D_i^j 识别为目标 $D_i^k\left(k=1,2,\cdots,n,k\neq j\right)$ 的错误判别。

设目标 Z^j 在判别域 Z_0^j 和 Z_1^j 上的概率密度函数分别为 $f\left(z\middle|H_0\right)$ 和 $f\left(z^j\middle|H_1^j\right)$，则有

虚警率：$$P_f=\sum_{j=1}^{n}P\left(D_1^j\middle|H_0^j\right)=\sum_{j=1}^{n}\int_{Z_1^j}f\left(z^j\middle|H_0^j\right)dz \tag{19}$$

漏警率：$$P_m=\sum_{j=1}^{n}P\left(D_0^j\middle|H_1^j\right)=\sum_{j=1}^{n}\int_{Z_0^j}f\left(z^j\middle|H_1^j\right)dz \tag{20}$$

检测率：$$P_d=\sum_{j=1}^{n}P\left(D_1^j\middle|H_1^j\right)=\sum_{j=1}^{n}\int_{Z_1^j}f\left(z^j\middle|H_1^j\right)dz \tag{21}$$

误检率：$$P_e=\sum_{j=1}^{n}\sum_{k=1,j\neq k}^{n}P\left(D_1^j\middle|H_1^j\right)=\sum_{j=1}^{n}\sum_{k=1,j\neq k}^{n}\int_{Z_1^k}f\left(z^j\middle|H_1^j\right)dz \tag{22}$$

目标分类中，关心的是存在的目标的识别效果，识别率一般指检测率。根据定义可知，虚警率、检测率、漏警率与误检率之和为 1。在实际计算时，首先计算识别率，再计算误报率、漏报率，对于剩余系统识别出来的而实际不存在的目标种类作计数来计算分类的虚警率。对于多目标识别中的虚警率应该计算一定时间段内积累的虚警率。对于数据集，我们采用求平均的方式来计算整体的虚警率、漏警率、检测率、误检率。

深度学习通过误差的反向传播来调整神经网络权值，达到建模的目的。反向传播迭代次数从几万次逐步增加到数十万次，直到训练误差趋于收敛为止。最后通过计算模型在测试集上的平均准确率（average precision，*AP*）和所有类别的平均准确率均值（mean AP，*m AP*）来评价模型的好坏。*AP* 从召回率和准确率两个角度衡量检测算法的准确性。*AP* 是评价深度检测模型准确性最直观的标准，可以用来分析单个类别的检测效果。*mAP* 是各个类别 *AP* 的平均值，*mAP* 越高表示模型在全部类别中检测的综合性能越高[11]。

3.4　实验设计

首先将各个策略与 SSD512 进行单独结合进行相应的对比实验，表明各个策略的作用；然后将所有策略与 SSD512 结合，对最终的改进算法进行整体测评。

用训练集训练原始 SSD512，将此模型记为 M0，在 M0 基础上加入自适应阈值策略，生成模型 M1；在 M0 基础上加入动态局部区域放大策略，生成模型 M2；最后将 M0 与所有策略结合在一起，生成模型 M3。使用两数据库测试集对 M0，M1，M3 进行测试和对比。为突出低分辨率小目标检测效果，使用构造的小目标测试集分别对 M0 和 M2 进行测试和对比。

另外本文选取了 Faster R-CNN、不需要预训练模型的 DSOD300[15]（deeply supervised object detector）检测框架[26]和 YOLO 系列检测框架中的升级版 YOLOv2 544[16]，以及 SSD 的改进模型 DSSD[17]（deconvolutional single shot detector）作为深度学习对比算法，与 M3 对比 Web Dataset 和 KITTI 数据集上的检测效果。对比检测框架算法使用作者发布的官方代码中的默认参数设置，与 M3 在相同训练集中进行训练。利用 Web Dataset 和 KITTI 数据集中的测试集进行测试。

3.5　实验结果

实验结果见表 1、2，分别对比了模型 M0、M1、M3 在 KITTI 和 WD 数据集上普通测试集的识别与检测效果。

表 1　各模型识别精度对比

model	dataset	*AP*(%)			*mAP*(%)
		Person	Car	Cyclist	
M0	KITTI	73.36	71.53	65.32	70.07
	WD	71.59	69.63	62.75	67.99
M1	KITTI	77.18	72.35	68.69	72.74
	WD	73.52	70.45	64.83	69.61
M3	KITTI	87.42	86.73	84.38	86.18
	WD	82.92	76.34	72.63	77.31

表 2　各模型检测效果对比

model	dataset	P_f (%)	*Pm* (%)	*Pd* (%)	*Pe* (%)
M0	KITTI	20.21	19.34	41.32	19.13
	WD	19.25	21.38	38.83	20.54
M1	KITTI	12.31	13.29	57.84	16.56
	WD	15.17	14.49	52.45	17.89
M3	KITTI	6.33	8.69	73.45	11.53
	WD	9.24	10.19	70.16	10.41

对比表 1、2 中 M0 和 M3 测结果，在 KITTI 数据集中，各类目标检测的 *AP* 提高了 14~19%不等，*mAP* 提高了约 16.11%，虚警率降低 13.88%，检测率提高 32.13%，漏警率降低 10.65%，误检率降低 7.6%；在 WD 数据集中，各类目标检测的 *AP* 提高了 7~11%不等，*mAP* 提高了约 7.24%，虚警率降低 10.01%，检测率提高 31.33%，漏警率降低 11.19%，误检率降低 10.13%。各项指标提升明显，表明本文策略总体对于弥补 SSD512 缺陷的有效性。

对比表 1、2 中 M0 和 M1 检测结果，在 KITTI 数据集中，各类目标检测的 *AP* 提高了 1~4%不等，*mAP* 提高了约 2.67%，虚警率降低 7.90%，检测率提高 16.52%，漏警率降低 6.05%，误检率降低 2.57%；在 WD 数据集中，各类目标检测的 *AP* 提高了 1~3%不等，*mAP* 提高了约 1.62%，虚警率降低 4.08%，检测率提高 13.62%，漏警率降低 6.89%，误检率降低 2.65%。M1 模型是在 M0 基础上加入自适应阈值策略训练得到的，通过在两个数据库上的测试结果与 M0 对比我们可以发现，M1 相较于 M0，对多目标的检测率得到了较大提高，多目标检测的虚警率和漏警率降低明显，表明自适应阈值策略发挥了区分低置信度真目标和高置信度假目标的作用，能够有效降低 SSD512 对多目标检测的漏警率和虚警率。

表 3、4 对比了模型 M0、M2 在 KITTI 和 WD 数据集上低分辨率小目标测试集的检测效果。

表 3　M0 和 M2 模型低分辨率小目标识别精度

model	dataset	AP(%)			mAP(%)
		Person	Car	Cyclist	
M0	KITTI	13.63	19.38	9.73	14.25
	WD	8.59	16.33	8.53	11.15
M2	KITTI	77.45	80.19	58.68	72.11
	WD	65.62	70.49	52.37	62.83

表 4　M0 和 M2 模型低分辨率小目标检测效果

model	dataset	P_f (%)	Pm (%)	Pd (%)	Pe (%)
M0	KITTI	14.25	33.12	29.43	10.14
	WD	11.15	34.15	30.48	6.45
M2	KITTI	10.82	10.17	60.48	18.53
	WD	11.91	14.85	52.03	21.21

对比表 3、4 中 M0 和 M2 检测结果，在 KITTI 数据集中，各类目标检测的 *AP* 提高了 49~64%不等，*mAP* 提高了约 57.86%，虚警率降低 22.3%，检测率提高 50.34%，漏警率降低 19.26%，误检率降低 8.78%；在 WD 数据集中，各类目标检测的 *AP* 提高了 44~57%不等，*mAP* 提高了约 51.68%，虚警率降低 22.24%，检测率提高 45.58%，漏警率降低 15.63%，误检率降低 6.71%。M2 模型是在 M0 基础上加入加入动态局部区域放大策略训练得到的，通过在两个数据库上低分辨率小目标测试集的测试结果对比我们可以发现，M2 相较于 M0，对多目标低分辨率小目标的识别精度和检测率得到了较大提高，检测的误检率、虚警率、漏警率降低明显，表明动态局部区域放大策略对低分辨率小目标检测和识别的有效性。由于低分辨率弱小目标类别难以判定，M2 的错误检测多为分类错误造成的即误检率高，而 M0 多目标检测率极低，表明了 SSD512 深度卷积网络逐层抽取特征的同时导致低分辨率弱小目标信息丢失严重。

图 5 验证了 M3 模型中 R-Net 增益效果评估的有效性。第一行蓝色字体数字指示红色框是目标的置信度。c 表示粗检测器检测结果，F 表示精检测检测器检测结果。红色字体表示 R-net 的精度增益。正值和负值标准化为[0,1]和[-1,0)。通过对比可以发现对于粗略检测足够好或者优于精细检测的区域，R-net 给出较低的精度增益分数（第 1 列和第 2 列），并且对于精细检测比粗略检测好得多的区域(第 3 列)，R-net 给出较高的精度增益分数。

利用 Web Dataset 和 KITTI 数据集中的普通测试集进行测试。检测识别效果如表 5 所示，其中 *FPS* 代表算法运行的速度，帧率。

对比表 5 中 M3 和 其他深度学习对比算法检测结果，在 KITTI 数据集中，各类目标识别的 *AP* 提高了 4~16%不等，*mAP* 提高了约 9~15%不等，检测率提高 13~28%；在 WD 数据集中，各类目标识别的 *AP* 提高了 5~12%不等，*mAP* 提高了约 4~9%不等，检测率提高 10~34%。虽然检测识别速率比不上 DSOD300、DSSD513、YOLOv2 544 等检测算法，但是 *FPS* 也能达到 38 帧/s，能够满足实时性的要求。



图 5　R-Net 放大精度增益效果

表 5　各检测算法检测识别效果对比

method	dataset	AP(%)			mAP(%)	Pd (%)	FPS
		person	car	cyclist			
Faster R-CNN	KITTI	83.26	74.13	75.42	77.61	45.22	13.15
	WD	81.49	71.33	68.65	73.82	36.63	11.64
DSOD300	KITTI	77.43	72.26	68.38	72.69	58.68	58.23
	WD	70.73	69.39	67.04	69.05	52.32	50.35
DSSD513	KITTI	75.46	69.53	68.34	71.11	59.42	46.34
	WD	72.19	68.83	66.45	69.16	49.79	39.38
YOLOv2 544	KITTI	79.43	71.25	67.32	72.66	60.82	56.74
	WD	73.29	69.63	68.85	70.59	54.86	49.28
M3	KITTI	87.42	86.73	84.38	86.18	73.45	37.56
	WD	82.92	76.34	72.63	77.31	70.16	32.83

4　结束语

针对现有基于大数据和深度学习的目标检测框架对于高分辨率复杂大场景中低分辨率小目标识别效果较差，多目标检测的精度和实时性难以平衡的问题，改进了基于深度学习的目标检测框架 SSD，提出一种改进的多目标检测框架 DRZ-SSD，将其专用于复杂大交通场景多目标检测。经过实验验证，改进策略有效弥补了传统 SSD 的缺陷，在应对弱小目标、多目标、杂乱背景、遮挡等检测难度较大的情况时，均能获得较好的效果，实现了算法精度与运行速率的平衡。由于卷积神经网络的结构不适合处理时序信息，结合递归神经网络[21]（一类具有记忆功能的神经网络）来解决视频目标检测和跟踪问题，将是下一步工作的重点。

参考文献：

[1] 迟晓君，孟庆春，陈鹏. 基于最小风险的 Bayes 决策方法在交通检测中的应用 [J]. 计算机应用研究，2005，22 (12): 204-205. (Chi Xiaojun,

Meng Qingchun, Chen Peng. Application of Bayesian decision-making method based on minimum risk in traffic detection [J]. Application Research of Computers, 2005, 22 (12): 204-205.)

[2] 于凯，贾磊，陈宇强. 深度学习的昨天，今天和明天 [J]. 计算机研究与发展，2013, 50 (9): 1799-1804. (Yu Kai, Jia Lei, Chen Yuqiang. Yesterday, Today and Tomorrow of Deep Learning [J]. Computer Research and Development, 2013, 50 (9): 1799-1804.)

[3] Liu Wei, *et al*. SSD: Single Shot MultiBox Detector [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[4] Ren Shaoqing, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017, 39 (6): 1137-1149.

[5] Joseph R, , *et al*. You only look once: unified, real-time object detection [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 779-788.

[6] Lin Tsuang Yi, *et al*. Feature Pyramid Networks for Object Detection [J]. arXiv preprint arXiv: 1612. 03144, 2016.

[7] Gao Mingfei, *et al*. Dynamic Zoom-in Network for Fast Object Detection in Large Images [J]. arXiv preprint arXiv: 1711. 05187, 2017.

[8] Glen Berseth, Cheng Xie, *et al*. Progressive Reinforcement Learning with Distillation for Multi-Skilled Motion Control [J]. arXiv preprint arXiv: 1802. 04765, 2018.

[9] 陈前斌，何小强，吴攀，等. 基于部分可测马尔科夫决策过程业务感知的微基站休眠时长确定策略 [J]. 电子与信息学报，2018, 40 (1): 130-136. (Chen Qianbin, He Xiaoqiang, Wu Pan, *et al*. A strategy for determining the sleep duration of micro base stations based on service awareness in partially measurable Markov decision processes [J]. Journal of Electronics and Information Technology, 2018, 40 (1): 130-136.)

[10] 赵鹏，李大寨，王韬. 基于 Logistic 回归的零件图像区域提取 [J]. 计算机应用研究，2017, 34 (4): 1265-1268. (Zhao Peng, Li Dazhai, Wang Wei. Part Image Region Extraction Based on Logistic Regression [J]. Application Research of Computers, 2017, 34 (4): 1265-1268.)

[11] 冯小雨，梅 卫，胡大帅. 基于改进 Faster R-CNN 的空中目标检测 [J]. 光学学报，2018, 38 (6): 1-16. (Feng Xiaoyu, Mei Wei, Hu Dashuai. Aerial target detection based on improved faster R-CNN [J]. Acta Optica Sinica, 2018, 38 (6): 1-16.)

[12] 陈果，左洪福. 图像的自适应模糊阈值分割法 [J]. 自动化学报，2003, 29 (5): 791-796. (Chen Guo, Zuo Hongfu. Adaptive fuzzy threshold segmentation method for images [J]. Acta Automatica Sinica, 2003, 29 (5): 791-796.)

[13] 胡聪，屈瑾瑾，许川佩，等. 基于自适应池化的神经网络的服装图像识别 [J//OL]. 计算机应用，2018, 1-8. (Hu Cong, Qu Wei, Xu Chuanpei, Zhu Aijun. Apparel image recognition based on adaptive pooling neural network [J//OL]. Computer application, 2018, 1-8.)

[14] 马春庭，郑坚，陈东根，等. 地面战场侦察系统多目标识别的评价指标 [J]. 探测与控制学报, 2006, 28 (1): 6-9. (Ma Chunting, Zheng Jian, Chen Donggen, *et al*. Evaluation index of multi-target recognition for ground battlefield reconnaissance system [J]. Journal of Detection & Control, 2006, 28 (1): 6-9.)

[15] Shen Zhiqiang, *et al*. DSOD: learning deeply supervised object detectors from scratch [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2017: 1937-1945.

[16] Zhang Jianming, *et al*. A real-time Chinese traffic sign detection algorithm based on modified YOLOv2 [J]. Algorithms, 2017, 10 (4): 127.

[17] Fu Chengyang, Liu Wei, Ranga A *et al*. DSSD: deconvolutional single shot detector [J]. arXiv preprint arXiv: 1705. 09587, 2017.

[18] 周飞燕，金林鹏，董军. 卷积神经网络研究综述 [J]. 计算机学报，2017, 40 (6): 1229-1251. (Zhou Feiyan, Jin Linpeng, Dong Jun. A review of convolutional neural networks [J]. Chinese Journal of Computers, 2017, 40 (6): 1229-1251.)

[19] 常亮，邓小明，周明全，等. 图像理解中的卷积神经网络 [J]. 自动化学报，2016, 42 (9): 1300-1312. (Chang Liang, Deng Xiaoming, Zhou Mingquan, *et al*. Convolutional neural networks in image comprehension [J]. Acta Automatica Sinica, 2016, 42 (9): 1300-1312.)

[20] Tang Pengjie, Wang Hanli, Kwong S. G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition [J]. Neurocomputing, 2017, 225 (2): 188-197.

[21] Munaro Matteo, *et al*. OpenPTrack: Open source multi-camera calibration and people tracking for RGB-D camera networks [J]. Robotics & Autonomous Systems, 2016, 75: 525-538.